

Contents

1	Introduction	1
2	Time complexity	3
2.1	NP	3
2.2	E	3
3	Powers and complexity	5
3.1	Square-free words	6
3.2	Cube-free words	7
3.3	Strongly cube-free words	7
3.4	Almost square-free words	10
4	Appendix	12
4.1	Proof of Theorem 16	12
4.2	Proof of Lemma 14	12
4.3	Extended Pigeonhole Principle	12
4.4	Proof of Lemma 11.	13
4.5	Proof of Theorem 32	13
4.6	Proof of Lemma 23	15
4.7	Gelfond on arithmetic progressions	16
4.8	Proof that the word w in Theorem 24 is strongly cube-free. .	16
4.9	An illustration	17

Nondeterministic automatic complexity of almost square-free and strongly cube-free words

Kayleigh K. Hyde
Bjørn Kjos-Hanssen

February 18, 2014

Abstract

Shallit and Wang studied deterministic automatic complexity of words. They showed that the automatic Hausdorff dimension $I(\mathbf{t})$ of the infinite Thue word satisfies $1/3 \leq I(\mathbf{t}) \leq 2/3$. We improve that result by showing that $I(\mathbf{t}) \geq 1/2$. For nondeterministic automatic complexity we show $I(\mathbf{t}) = 1/2$. We prove that such complexity A_N of a word x of length n satisfies $A_N(x) \leq b(n) := \lfloor n/2 \rfloor + 1$. This enables us to define the complexity deficiency $D(x) = b(n) - A_N(x)$. If x is square-free then $D(x) = 0$. If x almost square-free in the sense of Fraenkel and Simpson, or if x is a strongly cube-free binary word such as the infinite Thue word, then $D(x) \leq 1$. On the other hand, there is no constant upper bound on D for strongly cube-free words in a ternary alphabet, nor for cube-free words in a binary alphabet.

The decision problem whether $D(x) \geq d$ for given x , d belongs to $NP \cap E$.

1 Introduction

The Kolmogorov complexity of a finite word w is roughly speaking the length of the shortest description w^* of w in a fixed formal language. The description w^* can be thought of as an optimally compressed version of w . Motivated by the non-computability of Kolmogorov complexity, Shallit and Wang studied a deterministic finite automaton analogue.

Definition 1 (Shallit and Wang [5]). The *automatic complexity* of a finite binary string $x = x_1 \dots x_n$ is the least number $A_D(x)$ of states of a deterministic finite automaton M such that x is the only string of length n in the language accepted by M .

This complexity notion has two minor deficiencies:

1. Most of the relevant automata end up having a “dead state” whose sole purpose is to absorb any irrelevant or unacceptable transitions.
2. The complexity of a string can be changed by reversing it. For instance,

$$A_D(011100) = 4 < 5 = A_D(001110).$$

If we replace deterministic finite automata by nondeterministic ones, these deficiencies disappear. The NFA complexity turns out to have other pleasant properties, such as a sharp computable upper bound.

Technical ideas and results. In this paper we develop some of the properties of NFA complexity. As a corollary we get a strengthening of a result of Shallit and Wang on the complexity of the infinite Thue word \mathbf{t} . Moreover, viewed through an NFA lens we can, in a sense, characterize exactly the complexity of \mathbf{t} . A main technical idea is to extend Shallit and Wang’s Theorem 9 which said that not only do squares, cubes and higher powers of a word have low complexity, but a word completely free of such powers must conversely have high complexity. The way we strengthen their results is by considering a variation on square-freeness and cube-freeness, *strong cube-freeness*. This notion also goes by the names of *irreducibility* and *overlap-freeness* in the combinatorial literature. We also take up an idea from Shallit and Wang’s Theorem 8 and use it to show that the natural decision problem associated with NFA complexity is in $E = \text{DTIME}(2^{O(n)})$. This result is a theoretical complement to the practical fact that the NFA complexity can be computed reasonably fast; to see it in action, for strings

of length up to 23 one can view automaton witnesses and check complexity using the following URL format

`http://math.hawaii.edu/wordpress/bjoern/complexity-of-110101101/`

and check one's comprehension by playing a Complexity Guessing Game at

`http://math.hawaii.edu/wordpress/bjoern/software/web/complexity-guessing-game/`

Let us now define our central notion and get started on developing its properties.

Definition 2. The nondeterministic automatic complexity $A_N(w)$ of a word w is the minimum number of states of an NFA M , having no ϵ -transitions, accepting w such that there is only one accepting path in M of length $|w|$.

The minimum complexity $A_N(w) = 1$ is only achieved by words of the form a^n where a is a single letter.

Theorem 3 (Hyde [3]). *The nondeterministic automatic complexity $A_N(x)$ of a string x of length n satisfies*

$$A_N(x) \leq b(n) := \lfloor n/2 \rfloor + 1.$$

Proof sketch. If x has odd length, it suffices to carefully consider the automaton in Figure 1. If x has even length, a slightly modified automaton can be used. \square

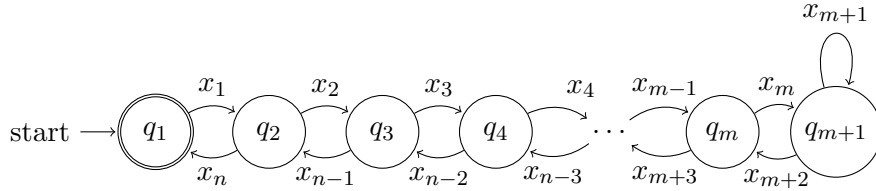


Figure 1: A nondeterministic finite automaton that only accepts one string $x = x_1x_2x_3x_4 \dots x_n$ of length $n = 2m + 1$.

Definition 4. The *complexity deficiency* of a word x of length n is

$$D_n(x) = D(x) = b(n) - A_N(x).$$

The notion of deficiency is motivated by the experimental observation that about half of all strings have deficiency 0; see Table 1.

Length n	$\mathbb{P}(D_n > 0)$	Length n	$\mathbb{P}(D_n > 0)$
0	0.000	1	0.000
2	0.500	3	0.250
4	0.500	5	0.250
6	0.531	7	0.234
8	0.617	9	0.207
10	0.664	11	0.317
12	0.600	13	0.295
14	0.687	15	0.297
16	0.657	17	0.342
18	0.658	19	0.330
20	0.641	21	0.303
22	0.633	23	0.322
24	0.593	25	0.283

(a) Even lengths.

(b) Odd lengths.

Table 1: Probability of strings of having positive complexity deficiency D_n , truncated to 3 decimal digits.

2 Time complexity

Definition 5. Let DEFICIENCY be the following decision problem.

Given a binary word w and an integer $d \geq 0$, is $D(w) > d$?

2.1 NP

Theorem 6 is not surprising; we do not know whether DEFICIENCY is NP-complete.

Theorem 6. DEFICIENCY *is in NP*.

Proof. Shallit and Wang’s Theorem 2 showed that one can efficiently determine whether a given DFA uniquely accepts w among string of length $|w|$. Hyde [3], Theorem 2.2, extended that result to NFAs, from which the result easily follows. \square

2.2 E

Definition 7. Suppose M is an NFA with q states that uniquely accepts a word x of length n . Throughout this paper we may assume that M contains

no edges except those traversed on input x . Consider the *almost unlabeled transition diagram* of M , which is a directed graph whose vertices are the states of M and whose edges correspond to transitions. Each edge is labeled with a 0 except for an edge entering the initial state as described below.

We define the *accepting path* P for x to be the sequence of $n + 1$ edges traversed in this graph, where we include as first element an edge labeled with the empty string ε that enters the initial state q_0 of M .

We define the *abbreviated accepting path* P' to be the sequence of edges obtained from P by considering each edge in order and deleting it if it has previously been traversed.

Lemma 8. *Let v be a vertex visited by an abbreviated accepting path $P' = (e_0, \dots, e_t)$. Then v is of one of the following five types.*

1. In-degree 1 (edge e_i), out-degree 1 (edge e_{i+1}).
2. In-degree 2 (edges e_i and e_j with $j > i$), out-degree 1 (e_{i+1}).
3. In-degree 1 (edge e_i), out-degree 2 (edges e_{i+1} and e_j , $j > i + 1$).
4. In-degree 2 (edges e_i and e_j with $j > i$), out-degree 2 (e_{i+1} and e_{j+1}).
5. In-degree 1 (edge e_t), out-degree 0.¹

Proof. The out-degree and in-degree of each vertex encountered along P' are both ≤ 2 , since failure of this would imply non-uniqueness of accepting path. Since all the edges of M are included in P , the list includes all the possible in-degree, out-degree combinations. We can define i by the rule that e_i is the first edge in P' entering v . Again, since all the edges of M are included in P , e_{i+1} must be one of the edges contributing to the out-degree of v , if any, and e_j must also be as specified in the types. \square

Lemma 8 implies that Definition 9 makes sense.

Definition 9. For $0 \leq i \leq t + 1$ and $0 \leq n \leq t + 1$ we let $E(i, n)$ be a string representing the edges (e_i, \dots, e_n) . The meaning of the symbols is as follows: 0 represents an edge. A left bracket $[$ represents a vertex that is the target of a backedge. A right bracket $]$ represents a backedge. The symbol $+$ represents a vertex of out-degree 2. When $i > n$, we set $E(i, n) = \varepsilon$. Next, assuming we have defined $E(j, m)$ for all m and all $j > i$, we can define $E(i, n)$ by considering the type of the vertex reached by the edge e_i . Let $a_i \in \{0, \varepsilon\}$ be the label of e_i .

¹This type was omitted by Shallit and Wang.

1. $E(i, n) := a_i E(i + 1, n)$.
2. $E(i, n) := a_i [E(i + 1, j - 1)] E(j + 1, n)$.
3. $E(i, n) := a_i + E(i + 1, n)$.
4. $E(i, n) := a_i [+E(i + 1, j - 1)] E(j + 1, n)$.
5. $E(i, n) := a_i E(i + 1, n)$.

Lemma 10. *The abbreviated accepting path P' can be reconstructed from $E(0, t)$.²*

Lemma 11.

$$|E(a, b)| \leq 2(b - a + 1).$$

Theorem 12. DEFICIENCY *is in* E .

Proof. Let w be a word of a length n , and let $d \geq 0$. To determine whether $D(w) > d$, we must determine whether there exists an NFA M with at most $\lfloor \frac{n}{2} \rfloor - d$ states which accepts w , and accepts no other word of length n . Since there are *prima facie* more than single-exponentially many automata to consider, we consider instead codes $E(0, t)$ as in Definition 9. By Lemma 10 we can recover the abbreviated accepting path P' and hence M from such a code. The number of edges t is bounded by the string length n , so by Lemma 11

$$|E(0, t)| \leq 2(t + 1) \leq 2(n + 1);$$

since there are four symbols this gives

$$4^{2(n+1)} = O(16^n)$$

many codes to consider. Finally, to check whether a given M accepts uniquely takes only polynomially many steps, as in Theorem 6. \square

Remark 13. The bound 16^n counts many automata that are not uniquely accepting; the actual number may be closer to 3^n based on computational evidence.

3 Powers and complexity

In this section we shall exhibit infinite words all of whose prefixes have complexity deficiency bounded by 1. We say that such a word has a hereditary deficiency bound of 1.

² Figure 2b in the Appendix gives an example of an automaton and the computation of $E(0, t)$.

3.1 Square-free words

Lemma 14. *Let a , b and \hat{a} be strings in an arbitrary alphabet with $ab = b\hat{a}$.*

- *Case 1: $|a| \leq |b|$. Then there is a string c and integers k and ℓ such that $a = \hat{a} = c^k$ and $b = c^\ell$.*
- *Case 2: $|a| \geq |b|$. Then there is a string u with $a = bu$ and $\hat{a} = ub$.*

In particular, if $a = \hat{a}$, then by symmetry we may assume that Case 1 obtains.

We will use the following simple strengthening from DFAs to NFAs of a fact used in Shallit and Wang's Theorem 9 [5].

Theorem 15. *If an NFA M uniquely accepts w of length n , and visits a state p as many as $k + 1$ times, where $k \geq 2$, during its computation on input w , then w contains a k th power.*

Proof. Let $w = w_0w_1 \cdots w_kw_{k+1}$ where w_i is the portion of w read between visits number i and $i + 1$ to the state p . Since one bit must be read in one unit of automaton time, $|w_i| \geq 1$ for each $1 \leq i \leq k$ (w_0 and/or w_{k+1} may be empty since the initial and/or final state of M may be p). For any permutation π on $1, \dots, k$, M accepts $w_0w_{\pi(1)} \cdots w_{\pi(k)}w_{k+1}$. Let $1 \leq j \leq k$ be such that w_j has minimal length and let $\hat{w}_j = w_1 \cdots w_{j-1}w_{j+1} \cdots w_k$. Then M also accepts

$$w_0w_j\hat{w}_jw_{k+1} \quad \text{and} \quad w_0\hat{w}_jw_jw_{k+1}.$$

By uniqueness,

$$w_0w_j\hat{w}_jw_{k+1} = w = w_0\hat{w}_jw_jw_{k+1}$$

and so

$$w_j\hat{w}_j = \hat{w}_jw_j$$

By Lemma 14, w_j and \hat{w}_j are both powers of a string c . Since $|\hat{w}_j| \geq (k - 1)|w_j|$, $w_j\hat{w}_j$ is at least a k th power of c , so w contains a k th power of c . \square

We next strengthen a particular case of Shallit and Wang's Theorem 9 to NFAs.

Theorem 16. *A square-free word has deficiency 0.*

Corollary 17. *There exists an infinite word of hereditary deficiency 0.*

Proof. There is an infinite square-free word over the alphabet $\{0, 1, 2\}$ as shown by Thue [7, 8]. The result follows from Theorem 16. \square

3.2 Cube-free words

Definition 18. For a word u , let $\text{first}(u)$ and $\text{last}(u)$ denote the first and last letters of u , respectively. A *weak cube* is a word of the form $u\text{first}(u)$ (or equivalently, $\text{last}(u)u$). A word w is *strongly cube-free* if it does not contain any weak cubes.

Theorem 19 (Shelton and Soni [6]). *The set of all numbers that occur as lengths of squares within strongly cube-free binary words is equal to*

$$\{2^a : a \geq 1\} \cup \{3 \cdot 2^a : a \geq 1\}.$$

Lemma 20. *If a cube www contains another cube xxx then either $|x| = |w|$, or $xx\text{first}(x)$ is contained in the first two consecutive occurrences of w , or $\text{last}(x)xx$ is contained in the last two occurrences of w .*

Theorem 21. *The deficiency of cube-free binary words is unbounded.*

Proof. Given k , we shall find a cube-free word x with $D(x) \geq k$. Pick a number n such that $2^n \geq 2k + 1$. By Theorem 19, there is a strongly cube-free binary word that contains a square of length 2^{n+1} ; equivalently, there is a strongly cube-free square of length 2^{n+1} . Thus, we may choose w of length $\ell = 2^n$ such that ww is strongly cube-free. Let $x = ww\hat{w}$ where \hat{w} is the proper prefix of w of length $|w| - 1$. By Lemma 20, x is cube-free. The complexity of x is at most $|w|$ as we can just make one loop of length w , with code (Theorem 12)

$$[w_1 \dots w_{\ell-1}]_{w_\ell}.$$

And so

$$\begin{aligned} D(x) &\geq \lfloor |x|/2 + 1 \rfloor - |w| \geq |x|/2 - |w| = \frac{3|w| - 1}{2} - |w| \\ &= |w|/2 - 1/2 \geq k. \end{aligned}$$

□

3.3 Strongly cube-free words

Theorem 22 (Thue [7, 8]). *The infinite Thue word*

$$\mathbf{t} = t_0 t_1 \dots = 0110 1001 1001 0110 \dots$$

given by

$$b = \sum b_i 2^i, \quad b_i \in \{0, 1\} \implies t_b = \sum b_i \pmod{2},$$

is strongly cube-free.

Lemma 23. For each $k \geq 1$ there is a sequence $x_{1,k}, \dots, x_{k,k}$ of positive integers such that

$$\sum_{i=1}^k a_i x_{i,k} = 2 \sum_{i=1}^k x_{i,k} \implies a_1 = \dots = a_k = 2$$

Let t_j denote bit j of the infinite Thue word. Then we can ensure that

1. $x_{i,k} + 1 < x_{i+1,k}$ and
2. $t_{x_{i,k}} \neq t_{x_{i+1,k}}$ for each $1 \leq i < k$.

Theorem 24. For an alphabet of size three, the complexity deficiency of strongly cube-free words is unbounded.

Proof. Let $d \geq 1$. We will show that there is a word w of deficiency $D(w) \geq d$. Let $k = 2d - 1$. For each $1 \leq i \leq k$ let $x_i = x_{k+1-i,k}$ where the $x_{j,k}$ are as in Lemma 23. Note that since $x_{i,k} + 1 < x_{i+1,k}$, we have $x_i > x_{i+1} + 1$. Let

$$\begin{aligned} w &= \left(2 \prod_{i=1}^{x_1-1} t_i \right)^2 t_{x_1} \left(2 \prod_{i=1}^{x_2-1} t_i \right)^2 t_{x_2} \left(2 \prod_{i=1}^{x_3-1} t_i \right)^2 \cdots t_{x_{k-1}} \left(2 \prod_{i=1}^{x_k-1} t_i \right)^2 \\ &= \lambda_1 t_{x_1} \lambda_2 \cdots t_{x_{k-1}} \lambda_k \end{aligned}$$

where $\lambda_i = (2\tau_i)^2$, $\tau_i = \prod_{j=1}^{x_i-1} t_j$, and where t_i is the i th bit of the infinite Thue word on $\{0, 1\}$, which is strongly cube-free (Theorem 22). Let M be the NFA with code (Theorem 12)

$$[+0^{x_1-1}]0[+0^{x_2-1}]0 \cdots 0 * [+0^{x_k-1}]$$

(where $*$ indicates the accept state). Let $X = \sum_{i=1}^k x_i$. Then M has $k-1+X$ many edges but only $q = X$ many states; and w has length

$$n = k - 1 + 2X = 2(d - 1) + 2X$$

giving $n/2 + 1 = d + X$.

Suppose v is a word accepted by M . Then M on input v goes through each loop of length x_i some number of times $a_i \geq 0$, where

$$k - 1 + \sum_{i=1}^k a_i x_i = |v|.$$

If additionally $|v| = |w|$, then by Lemma 23 we have $a_1 = a_2 = \dots = a_k$, and hence $v = w$. Thus

$$D(w) \geq \lfloor n/2 + 1 \rfloor - q = d + X - X = d.$$

In the Appendix we prove that w is strongly cube-free. □

Definition 2 yields the following lemma.

Lemma 25. *Let (q_0, q_1, \dots) be the sequence of states visited by an NFA M given an input word w . For any t, t_1, t_2 , and r_i, s_i with*

$$(p_1, r_1, \dots, r_{t-2}, p_2) = (q_{t_1}, \dots, q_{t_1+t})$$

and

$$(p_1, s_1, \dots, s_{t-2}, p_2) = (q_{t_2}, \dots, q_{t_2+t}),$$

we have $r_i = s_i$ for each i .

Note that in Lemma 25, it may very well be that $t_1 \neq t_2$.

Theorem 26. *Strongly cube-free binary words have deficiency bound 1.*

Proof. Suppose w is a word satisfying $D(w) \geq 2$ and consider the sequence of states visited in a witnessing computation. As in the proof of Theorem 32, either there is a state that is visited four times, and hence there is a cube in w , or there are three *state cubes* (states that are visited three times each), and hence there are three squares in w . By Theorem 19, a strongly cube-free binary word can only contain squares of length $2^a, 3 \cdot 2^a$, and hence can only contain powers u^i where $|u|$ is of the form $2^a, 3 \cdot 2^a$, and $i \leq 2$.

In particular, the length of one of the squares in the three state cubes must divide the length of another. So if these two state cubes are disjoint then the shorter one repeated can replace one occurrence of the longer one, contradicting Lemma 25.

So suppose we have two state cubes, at states p_1 and p_2 , that overlap. At p_1 then we read consecutive words ab that are powers $a = u^i, b = u^j$ of a word u , and since there are no cubes in w it must be that $i = j = 1$ and so actually $a = b$. And at p_2 we have words c, d that are powers of a word v and again the exponents are 1 and $c = d$.

The overlap means that in one of the two excursions of the same length starting and ending at p_1 , we visit p_2 . By uniqueness of the accepting path we then visit p_2 in both of these excursions. If we suppose the state cubes are chosen to be of minimal length then we only visit p_2 once in each excursion. If we write $a = rs$ where r is the word read when going from p_1 to p_2 , and s is the word going from p_2 to p_1 , then $c = sr$ and w contains $rsrsr$. In particular, w contains a weak cube. \square

Definition 27. For an infinite word \mathbf{u} define the *deterministic automatic Hausdorff dimension* of \mathbf{u} by

$$I(\mathbf{u}) = \liminf_{u \text{ prefix of } \mathbf{u}} A_D(u)/|u|.$$

and the *deterministic automatic packing dimension* of \mathbf{u} by³

$$S(\mathbf{u}) = \limsup_{u \text{ prefix of } \mathbf{u}} A_D(u)/|u|.$$

For nondeterministic complexity, in light of Theorem 3 it is natural to make the following definition.

Definition 28. Define the *nondeterministic automatic Hausdorff dimension* of \mathbf{u} by

$$I_N(\mathbf{u}) = \liminf_{u \text{ prefix of } \mathbf{u}} \frac{A_N(u)}{|u|/2}$$

and define S_N analogously.

Theorem 29 (Shallit and Wang's Theorem 18). $\frac{1}{3} \leq I(\mathbf{t}) \leq \frac{2}{3}$.

Here we strengthen Theorem 29.

Theorem 30. $I(\mathbf{t}) \geq \frac{1}{2}$. Moreover $I_N(\mathbf{t}) = S_N(\mathbf{t}) = 1$.

Proof. This follows from the observation that the proof of Theorem 26 applies equally for deterministic complexity. \square

3.4 Almost square-free words

Definition 31 (Fraenkel and Simpson [1]). A word all of whose contained squares belong to $\{00, 11, 0101\}$ is called *almost square-free*.

Theorem 32. A word that is almost square-free has a deficiency bound of 1.

Corollary 33. There is an infinite binary word having hereditary deficiency bound of 1.

Proof. We have two distinct proofs. On the one hand, Fraenkel and Simpson [1] show there is an infinite almost square-free binary word, and the result follows from Theorem 32. On the other hand, the infinite Thue word is strongly cube-free (Theorem 22) and the result follows from Theorem 26. \square

Conjecture 34. There is an infinite binary word having hereditary deficiency 0.

We have some numerical evidence for Conjecture 34, for instance there are 108 strings of length 18 with this property.

³ There is some connection with Hausdorff dimension and packing dimension. For instance, if the effective Hausdorff dimension of an infinite word \mathbf{x} is positive then so is its automatic Hausdorff dimension, by a Kolmogorov complexity calculation in Shallit and Wang's Theorem 9.

References

- [1] Aviezri S. Fraenkel and R. Jamie Simpson. How many squares must a binary sequence contain? *Electron. J. Combin.*, 2:Research Paper 2, approx. 9 pp. (electronic), 1995.
- [2] A. O. Gel'fond. Sur les nombres qui ont des propriétés additives et multiplicatives données. *Acta Arith.*, 13:259–265, 1967/1968.
- [3] Kayleigh Hyde. Nondeterministic finite state complexity. Master's thesis, University of Hawaii at Manoa, U.S.A., 2013.
- [4] Johannes F. Morgenbesser, Jeffrey Shallit, and Thomas Stoll. Thue-Morse at multiples of an integer. *J. Number Theory*, 131(8):1498–1512, 2011.
- [5] Jeffrey Shallit and Ming-Wei Wang. Automatic complexity of strings. *J. Autom. Lang. Comb.*, 6(4):537–554, 2001. 2nd Workshop on Descriptive Complexity of Automata, Grammars and Related Structures (London, ON, 2000).
- [6] R. O. Shelton and R. P. Soni. Chains and fixing blocks in irreducible binary sequences. *Discrete Math.*, 54(1):93–99, 1985.
- [7] A. Thue. Über unendliche zeichenreihen. *Norske Vid. Skrifter I Mat.-Nat. Kl., Christiania*, 7:1–22, 1906.
- [8] A. Thue. Über die gegenseitige lage gleicher teile gewisser zeichenreihen. *Norske Vid. Skrifter I Mat.-Nat. Kl., Christiania*, 1:1–67, 1912.

4 Appendix

4.1 Proof of Theorem 16

Proof of Theorem 16. Suppose w is a word of length $n = 2k$ or $n = 2k + 1$, of deficiency d . Then there is a witnessing automaton M with $q = k + 1 - d$ states. Since $n + 1 \geq 2k + 1 = 2(k + 1 - d) + 2d - 1 = 2q + (2d - 1)$, by the Extended Pigeonhole Principle (Theorem 35), there is a state p which is visited $2 + (2d - 1) = 3$ times $t_1 < t_2 < t_3$, during the $n + 1$ many times of the computation of M on input w (and is not visited at any other times in the interval $[t_1, t_3]$). By Theorem 15, w contains a square. \square

4.2 Proof of Lemma 14

Proof of Lemma 14. First of all, a and b are both prefixes of $ab = b\hat{a}$, so the shorter one among a, b is a prefix of the other, and the shorter one among \hat{a}, b is a suffix of the other.

Suppose that $|a| \leq |b|$. So $b = ad$ for some string d . If d is the empty string then we may let $c = a = b$. Similarly, if a is the empty string then the result is trivial with $k = 0$. We proceed by induction on length. Note $|\hat{a}| = |a|$. If $\max\{|a|, |b|\} \leq 1$ then the result is clear. Otherwise $\max\{|a|, |d|\} < |b| = \max\{|a|, |b|\}$ and $ad = b = d\hat{a}$ so by the inductive hypothesis, $a = c^i$, $d = c^j$ for some c , and consequently $b = c^{i+j}$.

Suppose now $|b| \leq |a|$. Then $a = bu$ for some u , and $\hat{a} = ub$. \square

4.3 Extended Pigeonhole Principle

Theorem 35 (Extended Pigeonhole Principle). *If $aq + d$ pigeons are placed in q pigeonholes where $d > 0$, then it cannot be the case that all pigeonholes have at most a pigeons; in fact, either*

- *there is a pigeonhole with at least $a + d$ pigeons; or*
- *there is a pigeonhole with at least $a + d - 1$ pigeons, and another with $a + 1$ pigeons; or*
- *there is a pigeonhole with at least $a + d - 2$ pigeons, and another with $a + 2$ pigeons; or*
- *there is a pigeonhole with at least $a + d - 2$ pigeons, and two others with $a + 1$ pigeons; or*

- all pigeonholes have at most $a + d - 3$ pigeons (which is impossible if $a + d - 3 \leq a$ and $d > 0$)

Proof of Theorem 35. Consider the maximum number of pigeons in a pigeonhole m . If $m \geq a + d$ we are in Case 1. If $m = a + d - 1$, we consider all the other pigeons and pigeonholes; there are then $q - 1$ pigeonholes and $aq + d - (a + d - 1) = a(q - 1) + 1$ pigeons. By the plain Pigeonhole Principle, there is a pigeonhole with at least $a + 1$ pigeons. If $m = a + d - 2$, we repeat the argument, consider the maximum number of pigeons in a pigeonhole other than a given one with the maximum number of pigeons. \square

4.4 Proof of Lemma 11.

Proof of Lemma 11. The four rules are

1. $E(i, n) = a_i E(i + 1, n)$
2. $E(i, n) = a_i [E(i + 1, j - 1)]_{a_j} E(j + 1, n)$
3. $E(i, n) = a_i + E(i + 1, n)$
4. $E(i, n) = a_i [+E(i + 1, j - 1)]_{a_j} E(j + 1, n)$

So either

$$|E(i, n)| \leq 2 + |E(i + 1, n)|$$

or

$$|E(i, n)| \leq 4 + |E(i + 1, j - 1)| + |E(j + 1, n)|$$

So if by induction hypothesis $|E(a, b)| \leq 2(b - a + 1)$ then

$$|E(i, n)| \leq 2 + 2(n - i - 1 + 1) = 2(n - i + 1)$$

or

$$|E(i, n)| \leq 4 + 2(j - 1 - i - 1 + 1) + 2(n - j - 1 + 1) = 2(n - i + 1)$$

\square

4.5 Proof of Theorem 32

Proof of Theorem 32. It is easy to verify for words of length at most 3. Suppose now w has length at least 4. Suppose w is a word of a length $n \in \{2k, 2k + 1\}$ where $k \geq 2$, with deficiency at least 2. Then there are $q = k - 1 \geq 1$ states occupied at $n + 1$ times. So $n + 1 \in \{2k + 1, 2k + 2\} =$

$\{2q + 3, 2q + 4\}$ times. There are at least $2q + 3$ times and only q states, so by the Extended Pigeonhole Principle (Theorem 35), we are in one of the following Cases 1–3.

- Case 1. There is at least one state that is visited at least 5 times. Then by Theorem 15, w **contains a fourth power**.
- Case 2. There is at least one state p_1 that is visited at least 4 times and another state $p_2 \neq p_1$ that is visited at least 3 times. Then by Theorem 15, there is a cube xxx and a square yy in w . Since w **has no squares of length** > 4 , we must have $|xx| \leq 4$ and $|yy| \leq 4$ and hence $1 \leq |x| \leq 2$ and $1 \leq |y| \leq 2$. We next consider possible lengths of x and y .
 - Subcase $|x| = 2$. Say $x = ab$ where $|a| = |b| = 1$. If $a \neq b$ then $xxx \in \{101010, 010101\}$ so 1010 occurs in w , but w **does not contain 1010**; if $a = b$ then 0000 or 1111 occurs in w , contra assumption.
 - Subcase $|x| = 1, |y| = 2$: In this case, the xxx and yy occurrences must be disjoint, because the states in a yy occurrence are $p_2p_3p_2p_3p_2$ for some p_3 which must be disjoint from $p_1p_1p_1p_1$ when $p_1 \neq p_2$. But then we can replace these by $p_2p_3p_2p_3p_2p_3p_2$ and p_1p_1 , respectively, giving two distinct state sequences leading to acceptance, contradicting Lemma 25.
 - Subcase $|x| = 1, |y| = 1$: In this case again the occurrences of xxx and yy must be disjoint, since $p_1 \neq p_2$. We can replace p_1^4 and p_2^3 by p_1 and p_2^6 , respectively, again contradicting Lemma 25.
- Case 3. There are at least 3 states p_1, p_2, p_3 (all distinct) that are each visited at least 3 times. Then by Theorem 15, there are three squares u_iu_i at three distinct states $p_i, 1 \leq i \leq 3$. By assumption $|u_iu_i| \leq 4$ so $|u_i| \leq 2$.
 - Subcase 3.1. $|u_i| = |u_j| = 1$ for two values $1 \leq i < j \leq 3$. Then the argument is entirely analogous to that in Case 2.
 - Subcase 3.2 $|u_j| = |u_k| = 2$ for two values $1 \leq j < k \leq 3$.
 - * Subsubcase 3.2.1. If disjoint, we can replace u_j^2 by u_k^2 to get u_k^4 , again a **fourth power**, by the argument of Subcase 3.1.
 - * Subsubcase 3.2.2. If nondisjoint with full overlap then

$$p_j a_1 p_j a_2 p_j$$

and

$$p_k b_1 p_k b_2 p_k$$

become

$$p_j p_k p_j p_k p_j p_k$$

and immediately we get 10101 **or** 01010 **or a fourth power in** w ;

- * Subsubcase 3.2.3. If partial overlap only then $p_j a_1 p_j a_2 p_j$ and $p_k b_1 p_k b_2 p_k$ become, by Lemma 25, $p_j a p_j a p_j$ and $p_k b p_k b p_k$ and then

$$p_j a p_j p_k p_j p_k b p_k$$

By Lemma 25 again, this must be

$$p_j p_k p_j p_k p_j p_k p_j p_k = (p_j p_k)^4$$

and so the read word must be of the form $abababa$, giving **an occurrence of** 1010 **(if** $a \neq b$) **or of a 7th power (if** $a = b$) **in** w .

□

4.6 Proof of Lemma 23

Proof of Lemma 23. Let

$$x_{1,1} = 1.$$

Given $x_{1,k-1}, \dots, x_{k-1,k-1}$, we let $x_{i,k} = 3x_{i,k-1}$ for $i < k$ and $x_{k,k} = 3u_{k-1} + 2$ for a sufficiently large number u_{k-1} . Reducing the equation

$$\sum_{i=1}^k a_i x_{i,k} = 2 \sum_{i=1}^k x_{i,k}$$

modulo 3, we see that $a_k \equiv 2 \pmod{3}$. If $a_k \geq 5$ then

$$\begin{aligned} \sum_i a_i x_{i,k} &\geq 5x_{k,k} = 15u_{k-1} + 10 \\ &> 6 \sum_{i < k} x_{i,k-1} + 6u_{k-1} + 4 = 2 \sum_{i < k} x_{i,k} + 2(3u_{k-1} + 2) = 2 \sum_{i \leq k} x_{i,k}; \end{aligned}$$

provided

$$3u_{k-1} + 2 > 2 \sum_{i < k} x_{i,k-1}$$

so we conclude $a_k = 2$. Then we can cancel a_k , divide by three and reduce to the induction hypothesis.

Thus our numbers are

$$\begin{aligned} x_{1,2} &= 3, & x_{2,2} &= 3u_1 + 2, \\ x_{1,3} &= 3^2, & x_{2,3} &= 3(3u_1 + 2), & x_{3,3} &= 3u_2 + 2 \end{aligned}$$

and in general

$$x_{j,k} = 3^{k-j}(3u_{j-1} + 2)$$

To ensure (1) we just take u_{j-1} sufficiently big. To ensure (2), we apply Lemma 36. \square

4.7 Gelfond on arithmetic progressions

Lemma 36. *Fix j and k and let t_x denote the x th bit of the Thue word. The function*

$$f(u) = t_{x(u)-1} \quad \text{where} \quad x(u) = 3^{k-j}(3u + 2)$$

is eventually nonconstant.

Proof. Gelfond [2] showed that \mathbf{t} has no infinite arithmetic progressions (see also Morgenbesser, Shallit, Stoll [4]). \square

4.8 Proof that the word w in Theorem 24 is strongly cube-free.

Proof that the word w in Theorem 24 is strongly cube-free. Suppose a word uu is contained in w .

Proof that the number of 2s in uu is either 0 or 2. Let o_1, \dots, o_{2a} denote the occurrences of 2s in uu and suppose $a \geq 1$. Let $\delta_i = o_{i+1} - o_i$. Then the sequence $(\delta_1, \dots, \delta_a)$ is an interval in the sequence

$$(x_1 - 1, x_1, x_2 - 1, x_2, \dots, x_{k-1} - 1, x_{k-1}, x_k - 1).$$

Since $x_i > x_{i+1} + 1$, in particular $|x_i - x_{i+1}| > 1$ and so this sequence is injective, i.e., no two entries are the same. But $(o_1, \dots, o_a) = (o_{a+1} - |u|, \dots, o_{2a} - |u|)$. So $\delta_{a+1} = o_{a+2} - o_{a+1} = o_2 - o_1 = \delta_1$ which implies $a = 1$.

So either Case 1 or Case 2 below obtains. **Case 1: The number of 2s in uu is zero.** Then certainly $u\text{first}(u)$ is not contained in w , since the infinite Thue word is strongly cube-free. **Case 2: The number of 2s in uu is two.** Then we have one of the following two cases.

1. uu is contained in a word of the form

$$t_1 \cdots t_{x_i} \ 2 \ t_1 \cdots t_{x_{i+1}-1} \ 2 \ t_1 \cdots t_{x_{i+1}}.$$

We guard against that by making sure that

- $t_{x_i} \neq t_{x_{i+1}-1}$ (Lemma 23) and
- $2 \neq t_{x_{i+1}}$ (the Thue word uses only the letters 0 and 1)

2. uu is contained in a word of the form

$$t_1 \cdots t_{x_i-1} \ 2 \ t_1 \cdots t_{x_i} \ 2 \ t_1 \cdots t_{x_{i+1}-1}.$$

Since uu contains exactly two 2s and the t_j are not 2s, it follows that $uu = a2b2c$ where a, b, c are words in the binary alphabet $\{0, 1\}$. Then $u = a2b_1 = b_22c$ where $b = b_1b_2$, so $a = b_2$, $c = b_1$ and so actually $u = a2c$ and $t_1 \cdots t_{x_i} = b = ca$. Here then $|ca| = x_i$. If $|a| \leq 2$ then consequently

$$x_i - 2 \leq |c| \leq x_{i+1} - 1$$

which contradicts $x_{i+1} < x_i - 1$. If $|a| \geq 2$ then we appeal to Lemma 37.

□

Lemma 37. $t_{x_i-2}t_{x_i-1}2t_1 \cdots t_{x_i}2$ cannot be part of a square having only two 2s.

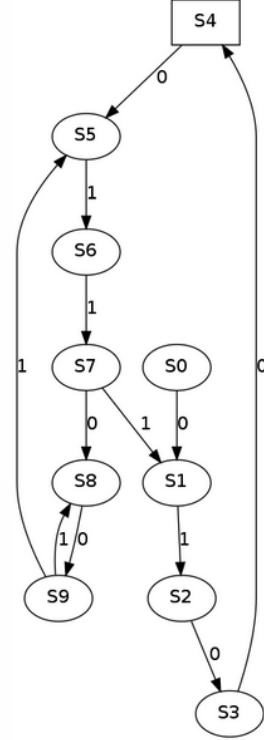
Proof. The Thue word is made up of disjoint occurrences of the words 01 and 10. Each of these two words are of the form $z\bar{z}$ where $\bar{z} = 1 - z$. The idea now is that if x_i is odd then say it ends in a lone 0 and 2, 02; then adding the next control bit will give something ending in 012, preventing a square.

More precisely, since $t_1 \cdots t_{x_i-1}2$ having odd or even length ends in say $z\bar{z}2$ or $z\bar{z}a2$ respectively, and then $t_1 \cdots t_{x_i-1}t_{x_i}2$ ends in $z\bar{z}b2$ or $z\bar{z}a\bar{a}2$, respectively; either way $t_1 \cdots t_{x_i-1}2$ and $t_1 \cdots t_{x_i-1}t_{x_i}2$ are incompatible. □

4.9 An illustration

$E(i, n)$	Computation
$E(0, 12)$	$\varepsilon E(1, 12) = E(1, 12)$
$E(1, 12)$	$a_1[E(2, 11)]_{a_{12}} E(13, 12)$
$E(13, 12)$	ε
$E(2, 11)$	$a_2 E(3, 11)$
$E(3, 11)$	$a_3 E(4, 11)$
$E(4, 11)$	$a_4 E(5, 11)$
$E(5, 11)$	$a_5[E(6, 10)]_{a_{11}} E(12, 11)$
$E(6, 10)$	$a_6 E(7, 10)$
$E(7, 10)$	$a_7 + E(8, 10)$
$E(8, 10)$	$a_8[E(9, 9)]_{a_{10}} E(11, 10)$
$E(9, 9)$	$a_9 + E(10, 9) = a_9 +$
$E(8, 10)$	$a_8[a_9+]_{a_{10}}$
$E(7, 10)$	$a_7 + a_8[a_9+]_{a_{10}}$
$E(6, 10)$	$a_6 a_7 + a_8[a_9+]_{a_{10}}$
$E(5, 11)$	$a_5[a_6 a_7 + a_8[a_9+]_{a_{10}}]_{a_{11}}$

(a) The + marks the place of a loopback. length $n = 22$.



(b) Complexity witness for the string 010001100101010111100, one of 2,655,140 simple strings of length $n = 22$.

Figure 2: The code is $E(0, 12) = a_1[a_2 a_3 a_4 a_5[a_6 a_7 + a_8[a_9+]_{a_{10}}]_{a_{11}}]_{a_{12}}$ where $(a_1, \dots, a_{12}) = (0, 1, 0, 0, 0, 1, 1, 0, 0, 1, 1, 1)$. In reduced form, $E(0, 12) = 0[0000[00 + 0[0+]]]$.